NASA-AISRP YEAR 1 ANNUAL REPORT
Report Period: 10/1/2004-9/30/2005
Grant No: NNG05GA30G

Estimating Missing Data in Sensor Network Databases Using Data Mining to Support Space Data Analysis

Le Gruenwald
University of Oklahoma
School of Computer Science
200 Felgar Street, Room 116 EL
Norman, OK 73019
Phone: 405-325-3498
Fax: 405-325-4044
Email: ggruenwald@ou.edu

1. PROJECT SUMMARY

Recent advances in Micro Electro Mechanical Systems (MEMS) based sensor technology, low-power analog and digital electronics, and low-power Radio Frequency (RF) design have made possible the development of relatively inexpensive and low-power wireless micro sensors that can be integrated in a network. The purpose of such a network is to monitor, combine, analyze and probably respond to the data collected by hundreds (or even thousands) sensors distributed in the physical world in a timely manner. This network can be used to support space data collection and analysis. For example, to facilitate solar system exploration missions, mobile sensors mounted on robots as well as hundreds of static micro sensors can be placed on MARS to collect its data and to send the collected data to a base station residing on MARS for real-time data analysis. The base station can then use the analysis results in real-time to determine actions that the robots should take. However, in a wireless sensor network, a significant amount of sensor readings sent from the sensors to the data processing point(s) (servers) may be lost or corrupted. In this research we propose a power-aware technique that uses association rules mining to handle such a problem. In this technique, to save battery power on sensors and to meet real-time requirements for data analysis, instead of requesting the sensor nodes (MS), the readings of which are missing, to resend their last readings, an estimation of the missing value(s) is performed by using the values available at the sensors relating to the MS through association rule mining. Temporal data mining using data clustering is also employed to improve data estimation. This research derives solutions for both centralized and distributed wireless sensor networks where transmissions can be single hops or multiple hops, and sensors/servers can be static or mobile. It then conducts performance evaluations using NASA sensor data to compare its proposed technique with existing statistical approaches.

2. PROJECT ACCOMPLISHMENTS

In Year 1 of the project, we were able to complete the following tasks:

- Conducted a literature survey of existing data mining techniques for data streams.
- Developed an algorithm called WARM (Window Association Rule Mining) to estimate missing sensor data for a single hop transmission sensor network.
- Conducted an investigation of existing statistical approaches for estimating missing data and identified the candidates to be compared with WARM.
- Implemented WARM and four selected statistical methods, Mean Substitution, Curve Estimation, Simple Linear Regression, and Multiple Linear Regression, and conducted experiments comparing them in terms of estimation accuracy, estimation response time, and memory space usage.
- Developed a framework for a data mining algorithm that deals with concept drifting when classifying data streams.

In the following sections, we provide details of the above tasks.

2.1. Conducted a literature survey of existing data mining techniques for data streams

Many recent applications, like sensor networks, network traffic monitoring, and e-commence click-stream, generate data streams that are extremely high volume, even unbounded, because they are continuously generated probably at a high speed. Recently techniques for mining data streams have been extensively researched. However, to the best of our knowledge, there has not a comprehensive survey of these techniques existing in the literature yet. Inspired by this, we surveyed available stream data mining techniques targeting the tasks of classification, clustering, and association rules analysis. We proposed taxonomy to classify the algorithms for each task. Also we proposed a feature-based model for each task and used it to study and compare these techniques. Although this is not an exhaustive survey, we believe that these surveyed techniques represent the current state of research on stream data mining. The results of the survey helped guide us in developing our data mining techniques to be used for estimating sensor data, which is a form of data streams and is the focus of our project.

2.2. Developed an algorithm called WARM (Window Association Rule Mining) to estimate missing sensor data for a single hop transmission sensor network

We assume that there is a single hop transmission sensor network where multiple sensors collect and send data to a central server for processing. In case that one of the sensors is not able to collect data or is not able to send the collected data to the server in time, the server, in response to a query that requests the missing data, will run an algorithm to estimate the missing data instead of waiting for the missing/late data to arrive. We have developed such an algorithm, called WARM (Window Association Rule Mining).

WARM first uses association rule mining to generate a set of sensors that are related to the sensor with the missing data (MS). It then uses the data generated by the related sensors to estimate the missing data, which is the data that the MS was not able to collect in time. The

related sensors' data contributes with different weights towards the estimated missing data depending on how many times each of those related sensors has yielded the same data as that of the MS in the past. If no related sensors can be found, WARM uses the average value of all data available in the current round of sensor data to estimate the missing data.

WARM makes use of the existing Apriori association rule mining technique [Agrawal, 1993], which has been developed to mine basket data, but modifies it to suit data stream mining. WARM considers association rules only with respect to a given state of data (e.g. low, medium, and high are three different states of temperature) in a sliding window of size *w*. Unlike Apriori which would identify all frequent itemsets before deriving possible association rules, to speed up the data estimation process, WARM only computes frequent 1-itemsets and frequent 2-itemsets. In addition, WARM includes three data structures: 1) a Buffer which stores the readings from the current round of sensor data; 2) a Cube which keeps track of all 1- and 2- itemsets observed in the last *w* rounds of sensor data; and 3) a Counter which stores the counters of all possible 1- and 2-itemsets. WARM consists of three algorithms: Update, CheckBuffer and Estimate. The Update algorithm updates both the Cube and the Counter with the data in the current round, which is stored in the Buffer. The CheckBuffer algorithm checks for missing values in the current round stored in the Buffer and initiates a proper action as a result of this check. The Estimate algorithm estimates a missing sensor reading using the data in the Buffer, Cube and Counter. The Estimate algorithm includes a weight assignment formula that computes the weight for each available sensor reading to contribute to the estimation of the missing sensor reading.

2.3. Conducted an investigation of existing statistical approaches for estimating missing data and identified the candidates to be compared with WARM

We have identified a number of popular statistical methods for data estimation existing in the literature [Allison 2002, Dempster, 1977, Gelman, 1995, Iannacchione, 1982, McLachlan, 1997, Rubin, 1987, Rubin, 1996, Shafer, 1995]. Some of them are listed below.

- *Mean Substitution*: replaces missing instances of a given variable with the mean value for that variable.
- *Simple Linear Regression*: replaces missing value by the value predicted from regression on observed variables.
- *Cold Deck Imputation*: replaces missing value by a value, which is independent of the data set.
- *Hot Deck Imputation*: replaces missing values with randomly selected values present in a pool of similar complete cases.
- *Expectation Maximization (EM)*: is a two-step iterative approach that estimates the parameters of a model starting from an initial guess. Each iteration consists of two steps: 1) an expectation step that finds the distribution for the missing data based on the known values for the observed variables and the current estimate of the parameters; and 2) a maximization step that substitutes the missing data with the expected value. The procedure iterates through these two steps until convergence is obtained. Convergence occurs when the change of the parameter estimates from iteration to iteration becomes negligible.

- *Maximum Likelihood*: uses all available data points in a database to construct the best possible first and second order moment estimates under the missing at random (MAR) assumption.
- *Multiple Imputations:* much like the EM algorithm, multiple imputations generates a maximum likelihood-based covariance matrix and vector of means and introduces statistical uncertainty into the model, and uses that uncertainty to replicate the natural variability found among the complete case data. It then imputes actual data values to fill in the incomplete data points in the data matrix, similar to the hot deck method. The difference is that it requires construction of five to ten databases with imputed values, each of which is analyzed individually. The results are then combined in one summary set of findings.

2.4. Implemented WARM and four selected statistical methods, Mean Substitution, Curve Estimation, Simple Linear Regression, and Multiple Linear Regression, and conducted experiments comparing them in terms of estimation accuracy, estimation response time, and memory space usage

We have written Java programs to implement and compare WARM with the four selected statistical methods. The experimental results showed the following: 1) WARM performs the best in terms of estimation accuracy (Table 1); 2) WARM performs the worse in terms of estimation response time (Table 2) and memory space usage (Table 3); and 3) The response time and memory space usage resulted from WARM, even though are the highest among those resulted from all the techniques studied, are within an acceptable range for many real-life sensor data applications. The results, thus, demonstrated that in general, WARM should be selected to estimate missing sensor data in single hop transmission networks.

Methods	RMSE (Root Mean Square Error)			
	winSize = 6	winSize = 18	winSize = 30	winSize = 42
Mean Substitution	0.141	0.146	0.146	0.144
Simple Linear Regression	0.247	0.246	0.238	0.237
Curve Estimation	0.252	0.26	0.245	0.247
Multiple Linear Regression	0.247	0.26	0.236	0.236
WARM	0.130	0.133	0.134	0.122

Table 1. Estimation Accuracy vs. Window Size

Methods	Response (Memory Access) Time (milliseconds)			
	winSize = 6	winSize = 18	winSize = 30	winSize = 42
Mean Substitution	6840	7560	8280	9000
Simple Linear Regression	7200	8640	10080	11400
Curve Estimation	7200	8640	10080	11400
Multiple Linear Regression	7560	9720	11880	14040
WARM	8575380	8658240	8730060	8822280

Table 2. Response Time vs. Window Size

Methods	Memory Space (KB)			
	winSize = 6	winSize = 18	winSize = 30	winSize = 42
Mean Substitution	1	2	3	4
Simple Linear Regression	1	2	3	4
Curve Estimation	1	2	3	4
Multiple Linear Regression	1	2	3	4
WARM	2670	4310	5950	7591

Table 3. Memory Space Usage vs. Window Size

2.5. Developed a framework for a data mining algorithm that deals with concept drifting when classifying data streams

Concept drifting occurs when the underlying data generating mechanism or the concept that we try to learn from the data is constantly evolving ([Yi, 2000, Hulten, 2001, Wang, 2003]. In order to solve the concept-drifting problem in stream data classification, this new algorithm will adaptively construct the model and use the following strategies:

- Quickly and efficiently detect concept drifting.
- Quickly and efficiently find the invalid nodes.
- Efficiently adapt the target model to future data if concept drifting occurs.
- Keep the current model for newly arriving records if no concept drifting occurs.

The overall idea of our algorithm is that by setting up an indicator for the current model, the algorithm monitors the indicator alone so that the concept drifting, if occurs, can be quickly detected. If the concept is stationary, the algorithm does not make any changes to the current model. If concept drifting is detected, the algorithm also uses an efficient searching algorithm to locate the invalid nodes. We are currently working on developing our framework further so that a complete data mining algorithm that can classify data streams and, at the same time, handles the concept drifting problem, can be achieved.

3. CHALLENGES

The major challenges that we faced in Year 1 are in the personnel category. Due to the arrival of the funding in the middle of the fall 2004 semester (October 2004), we were not able to recruit students to work on the project from its start date. Most of our students get their research assistant positions at the beginning of fall semesters and continue with the same positions for one or more years. Therefore, it was difficult for us to recruit students not only in fall 2004, but also in spring 2005.

4. PLANS FOR NEXT YEAR

In Year 2, we plan to work on the following tasks:

- Revise WARM to include temporal and spatial data mining.
- Work with NASA scientists to incorporate NASA data semantics into WARM.
- Conduct a theoretical analysis of WARM.
- Collect appropriate NASA data and conduct further performance evaluation comparing WARM with existing statistical methods.
- Extend WARM to include multi-hop transmission sensor networks.
- Conduct theoretical analysis and experiments comparing WARM with existing statistical methods for multi-hop transmission sensor networks using NASA data,

5. PUBLICATIONS AND PRESENTATIONS

1) Halatchev, Mihail and Le Gruenwald, "Estimating Missing Data in Related Sensor Data Streams", Proceedings of The International Conference on Management of Data, January 2005, pp. 83-94.
2) Le Gruenwald, "Estimating Missing Values in Related Sensor Data Streams", Presentation at Indian Institute of Technology (IIT), Delhi, India, January 2005.
3) Le Gruenwald, "Estimating Missing Values in Related Sensor Data Streams", Presentation at University of Paderborn, Germany, February 2005.
4) Le Gruenwald, "Estimating Missing Data in Sensor Network Databases Using Data Mining to Support Space Data Analysis" Presentation at NASA-AISRP Principal Investigator Meeting, NASA Ames Research Center, April 2005.
5) Le Gruenwald, "Estimating Missing Values in Related Sensor Data Streams", Presentation at University of Nanyang Technological University, Singapore, May 2005.
6) Halatchev, Mihail, Le Gruenwald and Nan Jiang, "A Window Association Rule Data Mining Technique to Estimate Missing Sensor Data", under preparation, to be submitted to Data and Knowledge Engineering Journal, July 2005.
7) Liu, Biao, Nan Jiang and Le Gruenwald, "A Survey of Stream Data Mining Techniques", under preparation, to be submitted to Journal of Very Large Data Bases, July 2005.
8) Jiang, Nan and Le Gruenwald, "Research Issues in Data Stream Association Rule Mining", under preparation, to be submitted to ACM SIGMOD RECORD, July 2005.

6. REFERENCES

[Agrawal, 1993] Rakesh Agrawal, Tomasz Imielinski, Arun Swami. “Mining Association Rules between Sets of Items in Large Databases”, the ACM SIGMOD International Conference on Management of Data, pp. 207-216, May 1993.

[Hulten, 2001] G. Hulten, L. Spencer, P. Domingos. Mining Time-Changing Data Streams, ACM International Conference on Knowledge Discovery and Data Mining, 2001, pp.97-106.

[Dempster, 1977] A. Dempster, N. Laird, and D. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society, Series B, 39(1), pages 1-38, 1977.

[Gelman, 1995] A. Gelman, J. Carlin, H. Stern, and D. Rubin. “Bayesian Data Analysis”. Chapman & Hall, 1995.

[McLachlan, 1997] G. McLachlan and K. Thriyambakam. “The EM Algorithm and Extensions”. New York: John Wiley & Sons, 1997.

[Rubin, 1987] D. Rubin, “Multiple Imputations for Nonresponce in Surveys”. New York: John Wiley & Sons, 1987.

[Rubin, 1996] D. Rubin. “Multiple Imputations after 18 Years”, Journal of the American Statistical Association, 91, pp. 473-478, 1996.

[Shafer, 1995] J. Shafer. “Model-Based Imputations of Census Short-Form Items”, The Annual Research Conference, Washington, DC: Bureau of the Census, pages *267-299*, 1995.

[Yi, 2000] B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, A. Biliris. On-Line Data Mining for Co-Evolving Time Sequences, International Conference on Data Engineering, 2000, pp. 13-22.

[Wang, 2003] Wang H., MiningConcept-Drifting Data Streams using Ensemble Classifiers. In 9th ACM International Conference on Knowledge Discovery and Data Mining SIGKDD, August 2003.